2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)

A Tool for Extracting Text from Scanned Documents and Convert it into Editable Format

S. Sukanya
PG Student
Department of Electronics and Communication Engineering
SSN College of Engineering
Chennai, India
sajisukanya@gmail.com

S. Joseph Gladwin
Associate Professor
Department of Electronics and Communication Engineering
SSN College of Engineering
Chennai, India
josephs@ssn.edu.in

C. Vinoth Kumar
Assistant Professor
Department of Electronics and Communication Engineering
SSN College of Engineering
Chennai, India
vinothkumarc@ssn.edu.in

***Abstract*— With the advent of Social media, now-a-days most of the data are stored in images. These data if processed correctly can provide large information. Hence there is a need to convert these data into information. A novel technique which can convert the data available in image into an editable format is proposed where the image can be acquired either by a camera, smart phone or directly from any source. The image is segmented into characters by using Connected Components (CC) and edge recombination using stroke width. This image is then converted to an editable format by using Optical Character Recognition (OCR) technology and Maximally Stable Extremal Region (MSER) for segmentation. The proposed system can also extract object from the images, which is done by using Artificial Neural Networks (ANN). The complete solution is developed using MATLAB and the output is stored in a variable. This can also be extracted to a word document if required where it can be edited. The performance of the system is measured by using two parameters namely precision rate and recall rate and has about 88% precision and 97% recall rate which is higher than most of the earlier proposed methods.**

Keywords— Optical Character Recognition; Maximally Extremal Region; Artificial Neural Networks; Connected Components.

I. INTRODUCTION

Optical Character Recognition (OCR) systems have been a subject of research long before the evolution of computers. With the increased development of technology, the OCR also have seen significant progress. It is known widely that as the more advanced the OCR system is, the more powerful the computer resources must be. The OCR technology combined with Intelligent Character Recognition (ICR) has changed the way the world handles documents. The most widely used field may be considered as business sector. OCR has enabled images to become more than just images but sources of information. With the use of this, people no longer need to type in the entire document, or manually search the document since all the process can be automated. The results obtained are accurate with less time than required when done manually.

The main challenge in text extraction is that the text in images may be of different size, style, and alignment. The images need not always be distortionless since many distortions may arise during the process of acquisition or due to the effects in camera. Sometimes the camera may not have a proper alignment with that of the scene to be captured. Thus the images may suffer from layout distortion. These images when subject to conversion may produce degraded results when compared to that obtained without any distortion. With the increase in the amount of distortion, the performance of the system also seems to reduce.

Since the system also has an Object feature extraction system, the security of many sensitive documents may be in peril. Hence a system which provides some security feature is required. Documents such as used in business deals with seals and signatures, hospital documents, attested documents, court orders, property papers, Government documents etc have to be protected from unlawful acts as they are more prone to misuse.

This paper is structured as follows. In section 2, the work related to the proposed work have been studied. Section 3 described the proposed work itself and the next section shows the results obtained. The following sections are conclusion and the references.

II. LITERATURE SURVEY

A. Object Extraction

Object Extraction can be done by using various techniques out of which the interactive one is the most famous. One method for object extraction is to acquire 3-D images and then convert these images into colored point cloud data and then to extract features from these colored point data [3]. Another method uses visual vocabulary and occurrence structure to detect pre-stored objects [5]. This method can hence detect only the trained objects and not more. The user scribbles are calculated and then the object boundary is calculated by taking those scribbles as shortest weighted path [6]. Other method for object detection than interactive method is to segment the image into regions [16]. Works on creating boundary boxes around objects have been proposed on a large scale but with very less accuracy [15]. Object detection was also proposed by using contour detection with an encoder-decoder network and with multiscale combinatorial grouping algorithm [4]. Object detection was also achieved by using Convolutional Neurals Networks [10]. Though many methods have been proposed, object recognition still remains to be an image dependent method.

B. Text Recognition.

One method for text extraction is to cluster character candidates into a tree [11]. This can be done by learned distance metric[18]. Text Recognition can also be done using numerous other parameters such as morphological operations. Automatic segmentation, extraction and recognition can be done using morphological technique

978-1-5386-9353-7/19/$31.00 ©2019 IEEE

which can be used even in complex images [9]. The layout and perspective distortion correction has also been widely studied and a variety of literature works are provided in this subject. Layout correction for Chinese text can be easily done due to the square nature of the characters [14]. Another method is to use the horizontal and vertical perspective shortening techniques [12]. Another method for text extraction is to analyse the connected components and structural analysis for grouping each edges into characters and later group each characters into words and text strings. A method for detecting text from images of documents are proposed using Markov Random Field [8]. Another similar work as that of proposed uses maximally stable extremal regions algorithm and geometric and stroke width transform are used to eliminate the false alarms [7]. Text detection is also done based on languages. One such work is done for Urdu language and the comparison of the performance OCR between Arabic, Urdu and English Language is also provided [13].

III. Proposed Work

The proposed system is developed using MATLAB programming and the complete system involves denoising, text extraction, object extraction, maintenance of log and Layout Distortion Correction technique. The denoising is done as part of pre-processing for the images. Figure 1 shows the overall flow of the proposed system. The system proposed recognises the text characters from images, extracts them and stores it into a variable. The value stored in the variable is the text content of the image which can also be extracted into a word document. Also any objects within the image can also be recognized and extracted along with the text, into the word document.

A. Work Flow

The proposed system converts a scanned document of any type such as jpg or png into an editable format. An image taken from any source can be given as input to the system for which it converts the computer recognizable ASCII character equivalent for the characters in the image. This is done with the help of OCR. The system also has provisions for correcting any distortions or noises present in the image. This is done with denoising and Layout Correction distortion technique.

The solution also confirm good performance in cases of presence of noise. The input image taken is initially subject to pre-processing techniques and distortion correction techniques. This corrected image is then given to segmentation where the image is segmented into character wise units for which MSER [2] is used. This character wise units are the input for the OCR engine which will compare each character to the pre-trained English alphabet and produces the most favorable results. The images of the documents may sometimes also include objects. A system which does not extract objects along with text may not provide much useful when most of the documents have certain parts to be treated as object. Hence this system also extracts the object contents from images and export in into the word document along with the text.

B. Optical Character Recognition

In 1990, during an effort to digitize historic papers was taken, the OCR technology became famous and came to widespread use. With time, the technology has been widely studied and developed for various purpose and applications. Before the advent of OCR, the only option available to digitize hardcopy was to manually re-type all the documents which proved to be a tedious and time-consuming process.

Optical Character Recognition is done in three steps- pre-processing, Character Recognition and post-processing. In pre-processing, the image is aligned in a vertical grid for separation of characters. This placement of the image in the grid is most important since the accuracy of the engine depends on this process. The vertical grid must not intersect with the black parts of the image which is considered to be the characters. The challenge here is to place the verticals grids between the characters and not in the middle of a character. The white space inside a character might be sometimes greater than the space between words. The next process is Character Recognition which recognizes characters from the segmented units. Each character from the grid is subject to feature extraction. The number of lines or the stroke width are taken as features and based on this, a classifier which can classify the detected character is built. Another method is pattern matching which will match the detected character to the already trained character set and gives the character which best matches the candidate character.

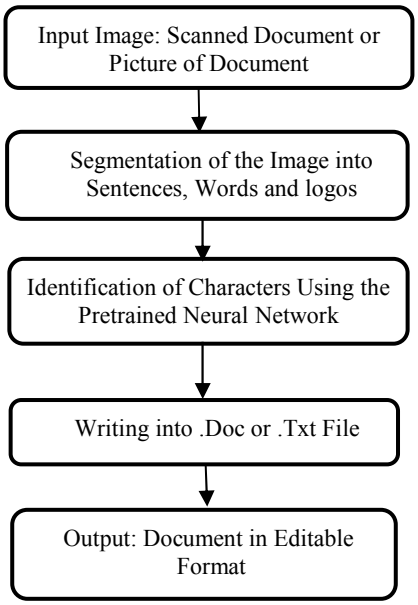


Fig 1 System Flow

C. Object Recognition

Documents containing objects are upmost importance and these objects also have to be treated and processed for extraction since the text content may sometimes prove to be meaningless without the objects. Hence a method to extract the objects have been proposed which uses bounding box and feature extraction for object detection.

The object extraction method proposed here is analogous to the one used in [1]. Figure 2 shows the method used for object extraction technique. The image is initially subject to Contour Edge Detection Network (CEDN) edge detection [17] and canny edge detection. Canny edge detection is used since it can detect even the finest contours. The results obtained from both these are combined together to get a

better result and bonding boxes are obtained along the candidate objects. A number of bounding boxes will be created around same object and hence non-maximal suppression is used to reduce the number of bounding boxes obtained thereby reducing redundancy. Saliency map is also performed to detect potential objects in the image. The texture features extracted include Contrast, Energy, Homogeneity and Entropy. Finally the detected image is extracted along with the text into a word document

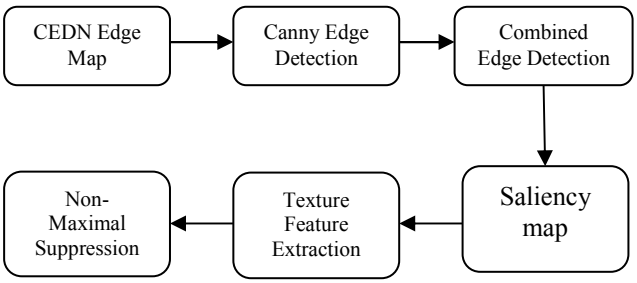


Fig 2 Object Extraction

D. Security

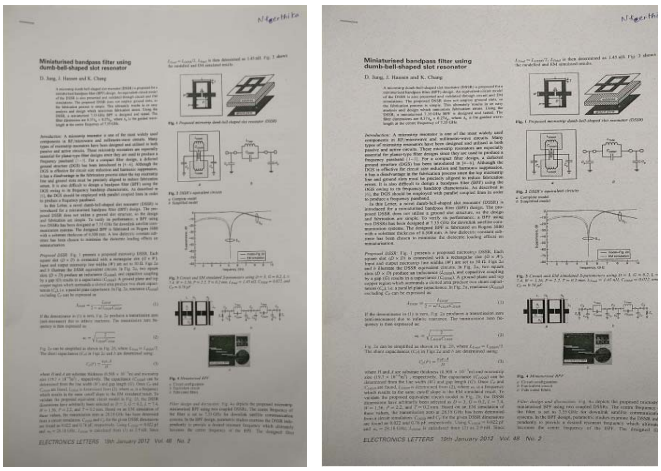
A system which includes object extraction clearly requires a security system since many document may have sensitive information misuse of which may lead to critical problems. Hence the system was developed along with security features wherein the user can type the username along with the password given to specific users thus allowing only authenticated persons to use the system. A log file was also maintained which will maintain the log of the users logged in to the system along with the date and time thus ensuring surveillance.

E. Denoising

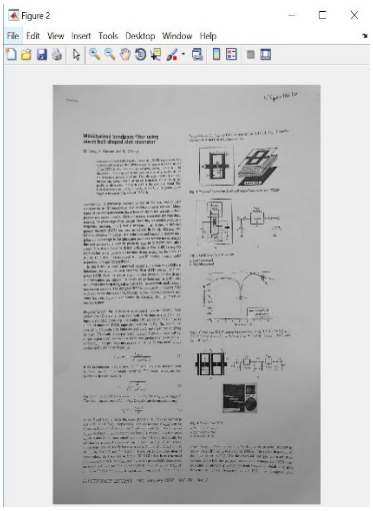
Denoising is considered to be image dependent and is essential that the type of image present in the image is to be known for applying the correct denoising algorithm for best accurate results. The noise we have concentrated is guassian noise since this is the most popular noise which occurs during acquisition and transferring. This can be reduced by using spatial filters such as guassian filter and median filter. Spatial filter calculates the spatial filter by finding the spatial filter from one point to a set of points. Figure 3(a) shows the input image that have been chosen, 3(b) shows the image added with noise and 3(c) shows the denoised image. Other noise types can also be addressed in this system by using the appropriate filter.

IV. Results and Discussions

After the text edit has been completed, the paper is ready for the template. The performance of the system seems to degrade severely with the presence of noise or distortion without any correction but after the correction techniques, the system provides considerable improvement in the performance. The accuracy of the system is measured with two parameter namely precision and recall rate. [10] claims to have around 79% precision and 69% recall rate whereas [19] claims to have 70% precision rate. 75 different images has been given as input to the system and the performance of the OCR engine was calculated. The proposed system shows a precision rate of 88% and 97% recall rate.



(a) (b)



(c)

Fig 3 Denoised Image

A. MATLAB GUI Output

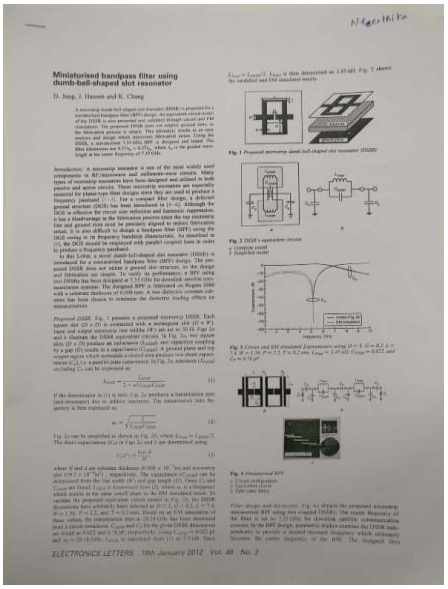


Fig 4 Input Image

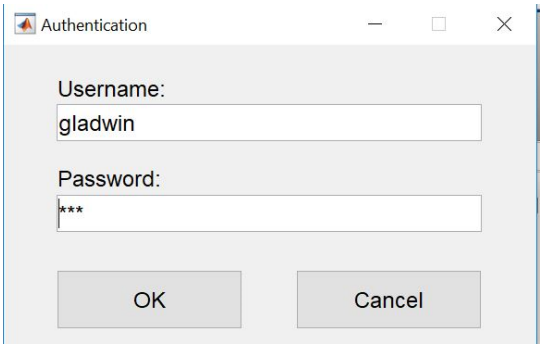


Fig 5 Authenticated Entry

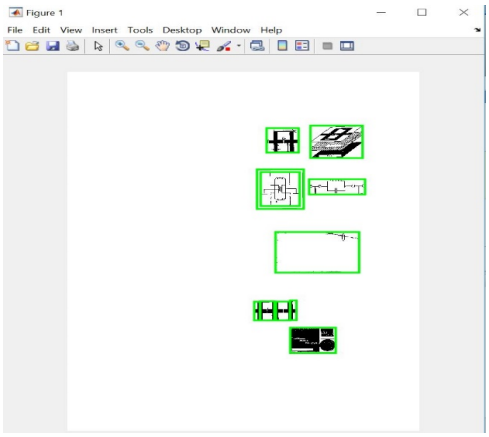


Fig 6 Object Recognition

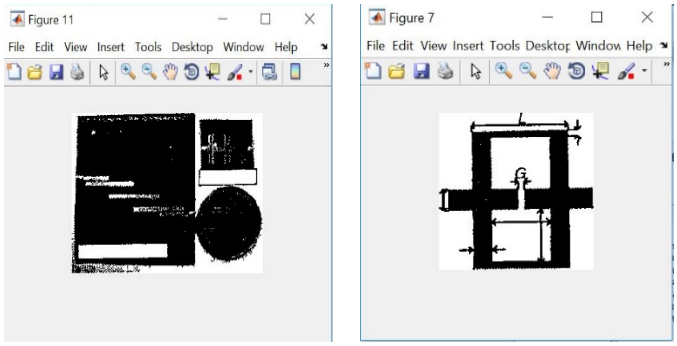


Fig 7 Object Extraction

The input image taken is showed in Figure 4. This image contains textual as well as object features both of which has to be extracted. The dialog box prompting for the username and password wherein the user will have to flourish the correct details for accessing the converter are shown in Figure 5. The recognised objects are enclosed in bounding boxes as shown in Figure 6 and these objects can also be extracted separately which is shown in Figure 7. Figure 8 shows the extracted text and object features extracted into a word document. The word document will be saved in the same directory as that of the current working drectory. Figure 9 shows the log file which is added as the second layer of security feature.

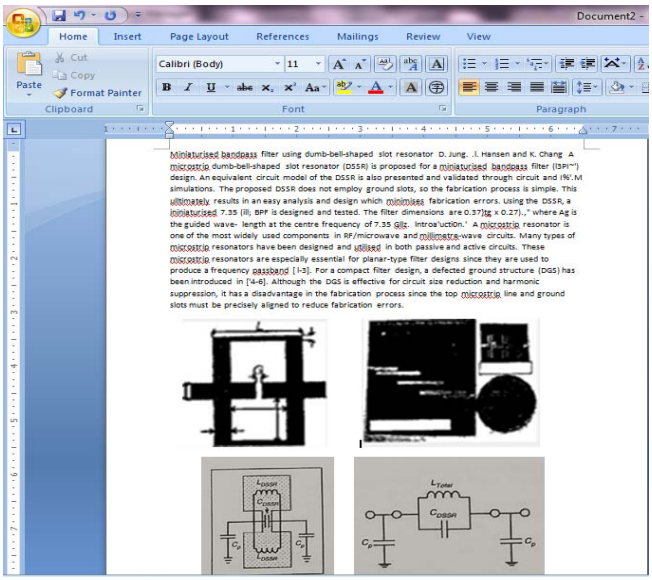


Fig 8 Exported Results

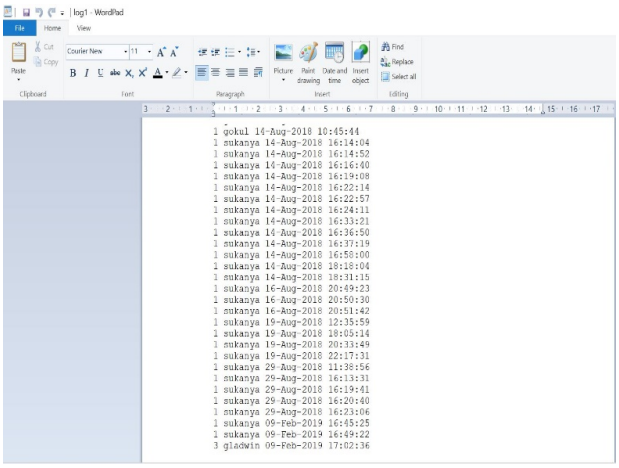


Fig 9 Log File

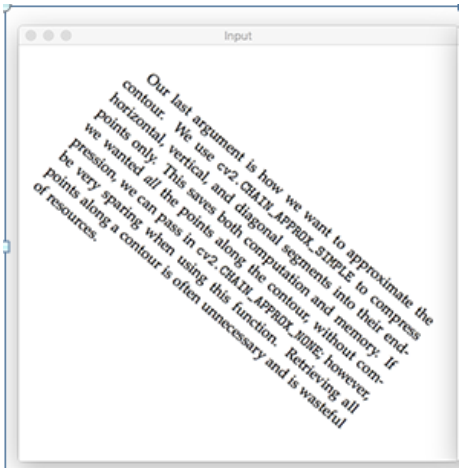


Fig 10 Example of Layout Distortion

B. *Layout Distortion Correction*

The images of documents cannot always be taken with high digital precision. Most of the times, the images of rare documents are taken under low lighting conditions and with high time-constraints. Hence the calibaration of the text in these images may not be qualitatively useful and a

correction mechanism seems vital. An example of a layout distorted image is given in Figure 10.

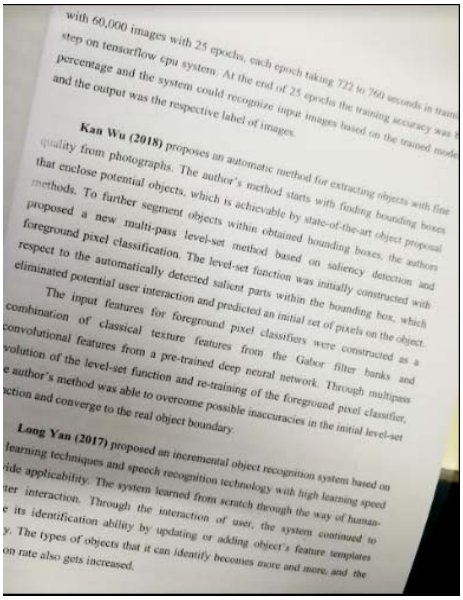


Fig 20 Input Image

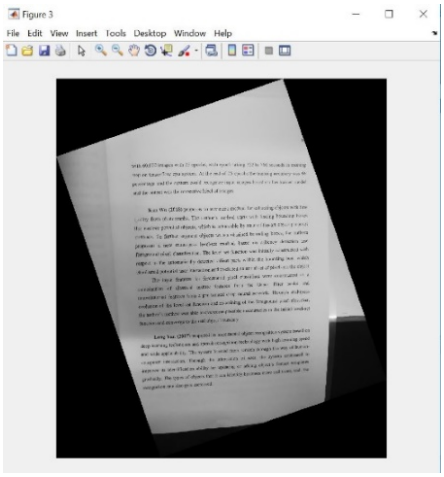


Fig 21 Distortion Corrected Image

V. CONCLUSION AND FUTURE WORK

Thus a technique which can convert data available in image into an editable format is proposed which is done with the help of OCR and can also extract objects from images. The proposed system produced best results for text and object extraction even under the presence of noise and distortion. The output are extracted to a word document where it can be edited and the security feature provide an add-on to the complete system.

The future work of this system may be to extend this to multiple languages and extraction of text features from video files. Dictionary may also be incorporated into the system for automatic grammatic accuracy and the system may be converted to a cloud based system such that an intelligent automatic system may be developed with remote access.

REFERENCES

[1] Kan Wu1 , Yizhou Yu, "Automatic object extraction from images using deep neural networks and the level-set method", The Institution of Engineering and Technology-2018.

[2] L. Gómez and D. Karatzas, ''MSER-based real-time text detection and tracking,''inProc.22ndInt.Conf.PatternRecognit.,2014,pp.3110–3115.

[3] Chi-Yi Tsai, Shu-Hsiang Tsai, "Simultaneous 3D Object Recognition and Pose Estimation Based on RGB-D Images", IEEE Access, vol. 6, pp. 28859-28869, 2018.

[4] Jimei Yang, Brian Price, Scott Cohen, et al, "Object Contour Detection with a Fully Convolutional Encoder-Decoder Network", IEEE Open Access, pp. 193-202, 2016.

[5] Vladimir Riffo , Domingo Mery, "Automated Detection of Threat Objects Using Adapted Implicit Shape Model", IEEE Transactions on Systems, Man, and Cybernetics Systems, pp . 472-482, 2016.

[6] Deselaers, D. Keysers, J. Hosang, and H. A. Rowley, "Gyropen: Gyroscopes for pen-input with mobile phones," IEEE Transactions on Human-Machine Systems, vol. 45, no. 2, pp. 263–271, 2015.

[7] Salahuddin Unar, Xingyuan Wang, Chuan Zhang, Chunpeng Wang, "Detected text-based image retrieval approach for textual images", IET Image Processing, vol. 13, 2019.

[8] Ismet Zeki Yalniz, R. Manmatha, "Dependence Models for Searching Text in Document Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 49-63, 2019.

[9] Vyankatesh V. Rampurkar, Sahil K. Shah, Gyankamal J. Chhajed, Sanjay Kumar Biswash, "An Approach towards Text Detection from Complex Images Using Morphological Techniques", Procc. of the Second International Conference on Inventive Systems and Control (ICISC 2018)

[10] Mr.Sudharshan Duth P, Ms.Swathi Raj, "Object Recognition in Images using Convolutional Neural Network", Proceedings of the Second International Conference on Inventive Systems and Control, 2018.

[11] X. Yin, Z. Zuo, S. Tian, and C. Liu, ''Text detection, tracking and recognition in video: A comprehensive survey,'' IEEE Trans. Image Process., vol. 25, no. 6, pp. 2752–2773, Jun. 2016

[12] C. Merino-Gracia, M. Mirmehdi, J. Sigut, "Fast Perspective Recovery of Text in Natural Scenes", Image and Vision Computing, vol.31,no.10,pp. 714-724, 2013.

[13] Naila Habib Khan, Awais Adnan, "Urdu Optical Character Recognition Systems: Present Contributions and Future Directions", IEEE Access, vol. 6, pp. 46019-46046, 2018.

[14] Yanwei Wang,Yuefang Sun, Changsong Liu, "Layout and Perspective Distortion Independent Recognition of Captured Chinese Document Image" , IAPR International Conference on Document Analysis and Recognition, 2014.

[15] P. Arbel´aez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, Multiscale combinatorial grouping. In CVPR, 2014.

[16] Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In CVPR, 2010.

[17] Yang J, Price Brian, Cohen S, et al., "Object contour detection with a fully convolutional encoder–decoder network'. Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp. 193–202, 2016.

[18] F. Yin and C.-L. Liu, "Handwritten Chinese text line segmentation by clustering with distance metric learning," Pattern Recognition vol. 42, no. 12, pp. 3146–3157, 2009

[19] Chong Yu, Yonghong Song, Quan Meng, Yuanlin Zhang, Yang Liu, "Text detection and recognition in natural scene with edge analysis", The institution of engineering and technology journal-2015.